Research Report

Unfair Treatment vs. Confirmation Bias? Comments on Santelices and Wilson

Neil J. Dorans

September 2010

ETS RR-10-20

Unfair Treatment vs. Confirmation Bias? Comments on Santelices and Wilson

Neil J. Dorans
ETS, Princeton, New Jersey

September 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Dan Eignor

Abstract

Santelices and Wilson (2010) claimed to have addressed technical criticisms of Freedle (2003) presented in Dorans (2004a) and elsewhere. Santelices and Wilson's abstract claimed that their study confirmed that SAT® verbal items do function differently for African American and White subgroups. In this commentary, I demonstrate that the Santelices and Wilson article provided no evidence to confirm differential item functioning (DIF) and failed to address my technical criticisms of Freedle. Instead, Santelices and Wilson contained several misrepresentations, including substituting "considered serious" for "more unusual" to describe an effect size and claiming to have studied four editions of the SAT when only two were actually studied. Central to its thesis was a hypothesis about DIF/difficulty correlations that was misattributed to Dorans. Santelices and Wilson based their argument for DIF on correlations between highly correlated variations on an index of DIF with item difficulty. While failing to demonstrate either unfairness in the test items or unfairness in the treatment of Freedle, Santelices and Wilson did contain evidence of confirmation bias.

Key words: confirmation bias, differential item functioning, differential prediction, score equity assessment, test fairness

Acknowledgments

The author thanks Brent Bridgeman, Daniel Eignor, Shelby Haberman, Michael Walker, Michael Zieky, and Rebecca Zwick for their helpful reviews of earlier versions of this commentary. The opinions expressed are those of the author. These opinions do not necessarily represent the position of ETS.

Freedle (2003) claimed that the SAT® was both culturally and statistically biased, and he proposed a solution to ameliorate this bias. Dorans (2004a) argued that these claims were based on serious analysis errors in Freedle. In particular, Dorans focused on how Table 2 in Freedle was constructed. The numbers in this critical Table 2 did not represent what Freedle claimed. When the table was constructed correctly, the effects that Freedle reported were reduced substantially in magnitude to the point where they did not warrant the corrective action, use of the R-score,[1] that Freedle proposed.

According to the abstract in Santelices and Wilson (2010, p. 106), Freedle (2003) "…faced an onslaught of criticism from experts at the Educational Testing Service (ETS)." A paper presented at a professional conference (Bridgeman & Burton, 2005), a commentary published in *Harvard Educational Review* (Dorans, 2004a), and two ETS research reports (Dorans & Zeller, 2004a, 2004b) constituted this "onslaught." The abstract later claims that Santelices and Wilson confirmed "…that SAT items do function differently for the African American and White subgroups in the verbal test and argue that the testing industry has an obligation to study this phenomenon" (p. 106). In this commentary, I demonstrate that Santelices and Wilson failed to address my technical criticisms of Freedle and failed to provide evidence of differential item functioning (DIF) on the test editions studied. Instead of addressing my concerns, the article misattributed to me the easy-to-refute hypothesis that choice of a DIF index appreciably affects the correlation between DIF and item difficulty.

In section 1, I briefly describe the purposes of DIF and distinguish it from other empirical fairness procedures. In section 2, I summarize pertinent aspects of the Santelices and Wilson (2010) article. In section 3, I restate my concerns with Freedle (2003). In section 4, I cite misrepresentations in Santelices and Wilson. In section 5, I report DIF and DIF/difficulty analysis based on all the data for the test edition that Santelices and Wilson emphasized. In section 6, I question the relevance of the DIF/difficulty correlation to the claim by Santelices and Wilson "...that SAT items do function differently for the African American and White subgroups" (p. 106). In section 7, I raise confirmation bias as a possible explanation for the design and execution of the Santelices and Wilson study.

1. Empirical Fairness Procedures

DIF has been used to screen SAT items since the late 1980s. On the SAT, the Mantel-Haenszel procedure is used to detect DIF, while the standardization procedure is used to describe DIF. Both procedures are described in Dorans and Holland (1993). In practice, the standardization approach supplements the Mantel-Haenszel method and is applied to responses to all item options, as well as to nonresponses (Dorans & Holland, pp. 50–57).

DIF is only one aspect of empirical fairness assessment (Dorans, 2004b). In a discussion of fairness, it is important to make a distinction between the test score and its constituent items. It is also important to make a distinction between the measurement of a construct and uses of a measure of that construct. Crossing *item vs. test* with *measurement vs. use* produces a 2-by-2 fairness framework, of which three cells are relevant to this discussion. DIF examines how well an item measures the construct of interest across subgroups. Score equity assessment (SEA) focuses on assessing whether the equivalence of scores from different editions of the same test holds across subgroups (Dorans & Liu, 2009). Differential prediction, based on the logically consistent fairness model (Petersen & Novick, 1976) attributed to Cleary (1968), focuses on whether test scores and other information predict criteria, such as grades in college, in much the same way across subgroups.

Both DIF and SEA assess fair measurement. DIF asks: Are items measuring what the total score measures in the same way across groups? SEA asks: Are different editions of the tests related to each other in the same way across groups? The fairness questions raised by Santelices and Wilson (2010) about access to higher education are score-use questions that cannot be addressed by a DIF analysis. Santelices and Wilson used the wrong diagnostic procedure.

Differential prediction addresses score use. These studies typically assess whether test scores, alone or with other information such as high school grades, predict first-year grade point averages equally well for different subgroups. Sackett, Borneman, and Connelly (2008) reported the well-established finding that SAT scores overpredict college performance for African American test-takers. In addition, Mattern, Patterson, Shaw, Kobrin, and Barbuti (2008) reported that high school grades overpredict grades in college more than SAT scores do. While use of both scores and grades reduces the overprediction, some differential prediction remains. See chapter 5 of Zwick (2002) for a discussion of hypotheses for these overpredictions.

2. Summary of Santelices and Wilson

Santelices and Wilson (2010) performed DIF and correlation analyses based on data from a nonrepresentative sample of students who took either one of what Santelices and Wilson claimed were two editions of the SAT administered in 1994 or one of two editions administered in 1999. Only students from California who indicated that English was their best language were included. The sample sizes in the Santelices and Wilson study were about 3,300 White and 700 African American test-takers for the 1994 editions and 6,600 White and 900 African American test-takers for the 1999 editions.

Santelices and Wilson (2010) studied nearly perfectly correlated variations on the standardization method for DIF assessment. They correlated these very highly related DIF measures with item difficulty, as measured by proportion-correct for all test-takers. As should be expected, they obtained essentially the same correlations between difficulty and DIF regardless of which standardization variation they used. Of the four "different" editions of the SAT, Santelices and Wilson focused their attention on the 1999 edition that had the largest correlation between DIF and difficulty.

Santelices and Wilson (2010) claimed to have confirmed Freedle's (2003) results by replicating the DIF/difficulty correlations across very highly related DIF indices and thereby demonstrated that the SAT is biased and invalid. Santelices and Wilson concluded by suggesting that the profession had dismissed Freedle's findings because of methodological criticisms by the ETS researchers, including myself, and that their study had dealt with these concerns.

3. My Concerns With Freedle's Calculations and R-Score

Instead of focusing on correlations as Santelices and Wilson (2010) suggested I did, my critiques (Dorans, 2004a; Dorans & Zeller, 2004a, 2004b) of Freedle (2003) focused on an unsound measurement practice advocated by Freedle and the miscalculations it was based on. Freedle recommended using two scores for reporting the performance of African American test takers on the SAT-Verbal: one score based on all the items and the other (called the R-score) based on questions that composed the harder half of the SAT-Verbal exam. Dorans (2004a) demonstrated that Freedle miscomputed performance on the hard half of the test, thereby inflating the influence of his R-score. Statistical support for the effect of reporting an R-score on test performance was greatly diminished by proper computation.

Dorans and Zeller (2004a) showed how the R-score differs from more conventional scores, such as number right or formula-score. They illustrated, via their Figures 1 and 2 and accompanying text, how inclusion of nonresponses in the calculation of percents correct on the hard half of the test translates into small R-score differences between African American and White test-takers. As explained in the next section, Freedle's (2003) large R-score effects were due to improper computation of proportions correct (percents correct divided by 100) on the hard half of the test.

Freedle (2003) used a variation of the standardization approach (Dorans & Holland, 1993) to DIF assessment. Standardization examines how items function in samples of test-takers from different subgroups who have the same total score. The most common standardization DIF index (STND P-DIF) can be viewed as a difference between the observed percent-correct in the *focal group,* such as African American test takers, and a "predicted" percent-correct that is obtained by summing, across score levels, the products of the conditional percents-correct at each score level in the *reference group,* such as White test-takers, by the relative number of African American test takers at each score level. The percent correct in the focal group is a measure of the item's difficulty in that group. The "predicted" percent-correct is what the item difficulty would be if the relationship in the focal group between success on the item as a function of total score was identical to the item/total score relationship in the reference group. If the observed and predicted percent-correct for the focal group are identical, item difficulty is unrelated to group membership, which indicates no DIF.

Freedle (2003) did not use STND P-DIF. Instead of computing the percent-correct (among those at a particular total score) as the number of people who answered the item correctly divided by all who were administered the item, he divided the number who answered correctly by the number who attempted the question. Whereas the denominator for a correctly calculated percent-correct is always the same (at a given score) across test questions, the denominator used by Freedle varies from question to question. Consequently, the "average" of these percents is not meaningful.

The kernel of Freedle's (2003) miscalculations can be summarized in a constructed example in which we examine proportions correct (percents answering the item correctly divided by 100) on two items from a long test that are achieved by two matched groups, where the groups are matched on number-correct score. Assume that one group (S) answers all questions

sequentially from beginning to end, while the other group (I) *iterates* through the test answering the easiest questions first, and then coming back to the hard questions. Let us assume there is no DIF on any item on the test, which means that for each question the proportions correct are equal in S and I. For example, let E represent a question that appears *early* in the test; the proportions correct for both the S and I groups are .8 on this item. On a question (L) that appears *late* in the test, the proportions correct are .4 in both I and S. The average proportion on the two items, E and L, in row 3 of numbers in Table 1, is .6 = (.8 + .4)/2. These numbers, .8, .6, and .4, appear in the total group columns of the table for the I (column 1 of *numbers*) and S (column 4 of *numbers*) groups, respectively.

Table 1

Illustration of Freedle (2003) and Correct Averaging

	Iterative group (I)			Sequential group (S)		
	Total I group (100%; 1)	Did answer item L (100%; 1)	Did *not* answer item L (0%; 0)	Total S group (100%; 1)	Did answer item L (50%; .5)	Did *not* answer item L (50%; .5)
Proportion-correct item E	.8	.8		.8	.8	.0
Proportion-correct item L	.4	.4		.4	.8	.0
Average proportion on E+L	$.6_c$	$.6_f$		$.6_c$	$.8_f$	.0
Freedle average	$.6_f$ = [(1 * .8) + (1 * .4)] / 2 = (.8 + .4)/2			$.8_f$ = [(1 * .8) + (1 * .8)] / 2 = (.8 + .8)/2		
Correct average	$.6_c$ = [(1 * .8) + (1 * .4)] / 2 = (.8 + .4)/2			$.6_c$ = {1 * .8 + [(.5 * .8) + (.5 * 0)]} / 2 = (.8 + .4 + 0)/2		

Note. Everyone (100%) in both groups S and I answered the early item (E), but 50% of S answered the late item, while 100% of I answered the item that appeared late in the test. The subscripts *f* and *c* stand for Freedle and Correct averaging.

All test-takers in both I and S answer item E. In addition, everyone in I answers item L; hence, the column 2 is equal to column 1, while column 3 is blank. Let's assume that only half of group S answers question L. Excluding the nonresponders (50% of the S group) from the calculation of the proportion correct for item L, as Freedle (2003) did, doubles the proportion-

correct from .4 (row 2, column 4) to .8 (row 2, column 5). Column 6 is the proportion-correct (.0) for the 50% who did not answer item L.

Summing the numbers in the *Did answer item L* column across the two items yields what appears to be a difference in performance between matched groups I and S in the *Average proportion* row. The S group has a larger "average," .8 (row 3, column 5), than the I group, .6 (row 3, column 2). Note that on question E, there are no nonresponders; the proportion in the *Did answer item L* group is the total group in both S and I groups. On question L, however, *only* those who answer are included in the calculation for both the I (100% responded) and S (50% responded) groups. This difference in nonresponse rates on question L gives the appearance that item L is easier for the S group (.8) than for the I group (.4).

In addition, the differential nonresponse rates across groups and items makes the average meaningless. For the S group, the average is based on the performance of all of S on item E and *half* of S on item L. For I, it is the average of performance of all of I on both item E and item L. As noted above, this is the kind of "averaging" that Freedle (2003) employed to create his Table 2. The calculations appear in the next to last row of Table 1. It gives the appearance (columns 2 and 5) that the average of items E and L favor the S group (.8) over the I group (.6).

Dorans (2004a) corrected Freedle's (2003) calculations by computing averages based on the *entire* group. These correct averages, .6 for this example, appear on the bottom row of Table 1. Since all in I responded to item L, the correct average in group I is equal to the Freedle average. In group S, however, the correct average is .2 lower than the Freedle average, because unlike the Freedle average, it uses the 50% of test-takers who obtained a 0 on item L in the calculation.

The effects of dropping the nonrespondents out of calculations of the hard half test averages in Freedle (2003) were complicated because of the number of items involved. This simple example, however, illustrates the effect that exclusion of nonresponders can have on calculations at the item and sum of items levels. When the expected averages on the hard half test were correctly computed in Dorans (2004a) with all the data including nonresponders, the large differences in hard half test performance between African American and White test-takers that Freedle used to justify his R-score were greatly reduced.

There is another major area of concern about Freedle's (2003) R-score suggestion. Dorans and Zeller (2004b) demonstrated that a hard half of the SAT-Verbal test could not

produce exchangeable scores with the full SAT test because of large differences in test difficulty and differences in reliability, with the hard half test producing less reliable scores for both African American and White test-takers. As a hard task becomes increasingly more difficult, subgroup differences become smaller, and the task becomes less and less useful for distinguishing among all but the ablest. For example, group differences in high jump performance between high school high jumpers and the rest of the student body would be zero if the high jump were set at 8 feet, which is just below the men's world record. In addition, we would have no information about how high the students could jump. Freedle observed that subgroup differences decreased as the bar was set higher on the SAT in the form of more difficult questions, postulated a reason for this reduction in differences, and suggested a solution, the R-score, which was based on the harder half of the test. The unintended consequence of Freedle's R-score suggestion was poorer measurement for all but the highest scoring test-takers. Dorans and Zeller (2004b, Table 4) reported a score reliability of .83 for African American test-takers on a 39-item hard half test in contrast to .89 on the 39-item easy half test. The hard half test is a sound measurement instrument for distinguishing among high-scoring test-takers, whether African American or White. It is a poor instrument for distinguishing among low-scoring test-takers, whether African American or White.

Santelices and Wilson (2010) failed to address either the miscalculations or the efficacy of the R-score. Instead, they misrepresented my work and their own data and misattributed an easily falsifiable hypothesis to me.

4. Misrepresentations

Santelices and Wilson (2010) contained several misrepresentations. The first misrepresentation involves word replacement and misattribution. In practice, SAT items are flagged for DIF on the basis of the Mantel-Haenszel procedure (Dorans & Holland, 1993, p. 42). On pages 49–50 of their paper, Dorans and Holland also provided effect sizes for the standardization method. Dorans and Holland labeled items exhibiting the largest effect size based on standardization as "more unusual." On page 116 of their paper, Santelices and Wilson reported effect size benchmarks for STND P-DIF, which are attributed to Dorans and Holland. Santelices and Wilson, however, substituted "considered serious" in place of "more unusual." In doing so, Santelices and Wilson created the false impression that "considered serious" was the

phrase Dorans and Holland used. To consider a STND P-DIF of .10 to be serious is a gross overstatement: 10 such items produce a raw score difference of 1 point. This misattribution and mischaracterization is a serious error, compounded by the extent to which the word *serious* was used throughout the Santelices and Wilson study.

In addition to misrepresenting Dorans and Holland (1993), Santelices and Wilson (2010) misrepresented their own data from their study. Santelices and Wilson claimed in their text and tables to have looked at four editions of the SAT. In endnote 17, however, Santelices and Wilson state that the two editions from the same year contained the same items. One edition was administered in November 1994 and the other one was administered in October 1999. Each of the two editions was administered in two different section orders.

Santelices and Wilson (2010) examined data from four samples of test-takers: two samples took the 1994 edition, and the other two took the 1999 edition. The -.41 and -.14 correlations between DIF and difficulty reported for the 1999 samples in Santelices and Wilson came from the same set of items. The fact that the largest and smallest correlations between DIF and difficulty across the four samples of test takers were associated with the *same set of items* is not mentioned in Santelices and Wilson. This is a serious omission. Inclusion of this fact in the article would have cast doubt upon the generalizability and stability of the correlations that they obtained, and the arguments based on those correlations.

The next misrepresentation found in Santelices and Wilson (2010) might be based on a lack of understanding of my critiques of Freedle (2003). It is central to their thesis, however. According to Santelices and Wilson, Dorans (2004a) in essence claimed that the correlations between DIF and difficulty reported by Freedle were an artifact of the standardization index that Freedle used. This is simply not true.

Contrary to what might be inferred from Santelices and Wilson (2010), I never disputed the existence of the correlation between DIF and item difficulty. I worked closely with Kulick and Hu (1989), who provided extensive documentation of this correlation. This correlation is an empirical fact that should remain fairly invariant across different highly related DIF indices, as a cursory examination of the standardization equations in Santelices and Wilson would reveal. (See the appendix to this commentary.) Contrary to expecting a zero correlation, there are good reasons to expect a nonzero correlation with real data.[2]

In fact, Dorans (2004a) made only one passing reference to the correlation between DIF and difficulty in the entire article. Instead of focusing on correlations as Santelices and Wilson (2010) suggested I did, my critiques (Dorans, 2004a; Dorans & Zeller, 2004a, 2004b) of Freedle (2003) focused on an unsound measurement practice (use of the R-score) advocated by Freedle, namely, reporting scores for individuals selected on the basis of their race that are based on the harder half of a test that is already very difficult for lower scoring test-takers, and the statistical justification provided for this suggestion.

In sum, Santelices and Wilson (2010) misrepresented Dorans and Holland (1993), misrepresented the data they used, and either misrepresented or misinterpreted my concerns with Freedle (2003). Instead of addressing the actual concerns, which were restated earlier in section 3 of this report, their study demonstrated that highly related DIF indices correlated with a measure of item difficulty to essentially the same degree.

In the next section, I present results based on complete data for the 1999 editions of the SAT that Santelices and Wilson (2010) focused on. Then in section 6, I question the relevance of the DIF/difficulty correlation to the fairness of the SAT score use in higher education.

5. DIF on the 1999 Test Edition Based on Complete Data

The variation observed in the Santelices and Wilson (2010) study for correlations between DIF and difficulty for the same items in two different samples of test takers illustrated how sensitive correlations between DIF and difficulty can be to sampling of test-takers, resulting in sampling error due in large part to the small sample sizes for the African American test-takers.[3] I decided to reduce some of this variability by examining the complete data on the 1999 test edition that they emphasized in their study.

Table 2 contains summary statistics for STND P-DIF and four measures of item difficulty for a nationwide sample containing 227,931 White and 28,753 African American test-takers administered either of the two section orders of the 1999 SAT test edition. The sampling error component of the DIF estimates based on this combined sample is about one sixth that of the Santelices and Wilson (2010) analysis.

Table 2

Summary Statistics and Correlations Among Standardized P-DIF and Three Measures of Item Difficulty Based on 78 Items From an SAT-Verbal Edition Administered in 1999

	STND P-DIF	$P_{a\text{-}a}^{+}$	Est $P_{a\text{-}a}^{+}$	P_{w}^{+}	P_{t}^{+}
Mean	0.00	0.49	0.49	0.62	0.60
SD	0.03	0.20	0.21	0.21	0.21
Max	0.06	0.89	0.90	0.96	0.94
Min	-0.09	0.10	0.09	0.14	0.14
	Correlations				
	STND P-DIF	$P_{a\text{-}a}^{+}$	Est $P_{a\text{-}a}^{+}$	P_{w}^{+}	P_{t}^{+}
STND P-DIF	1				
$P_{a\text{-}a}^{+}$	-0.18	1			
Est $P_{a\text{-}a}^{+}$	-0.32	0.99	1		
P_{w}^{+}	-0.30	0.98	0.98	1	
P_{t}^{+}	-0.29	0.98	0.99	1.00	1

The four measures of difficulty in Table 2 are proportion-correct in the African American test-takers ($P_{a\text{-}a}^{+}$); estimated proportion-correct for the African American test-takers (Est $P_{a\text{-}a}^{+}$); proportion-correct in the White test-takers (P_{w}^{+}); and proportion-correct in the total group of test-takers (P_{t}^{+}). The top half of the table contains means, standard deviations (SD), maxima, and minima for these statistics. The lower portion contains correlations among the difficulty indices and STND P-DIF.

A lack of DIF means that all the items function in essentially the same way as measures of the total score in both the African American and White test-taker samples. Mean DIF is typically close to zero; it is the standard deviation of DIF that matters. The first column in Table 2 contains a mean DIF of 0.0, an SD DIF of .03, a max DIF of .06, and min DIF of -.09. None of these items exhibited unusual amounts of DIF for African American and White test-takers on this edition of the SAT, nor were there any unusual amounts of DIF for items on the 1994 edition, which was also studied by Santelices and Wilson (2010).

These DIF results are not surprising. The SAT has screened pretest items for DIF since the late 1980s. As noted earlier, the Mantel-Haenszel approach is used for screening items. Items are classified as either A, B, or C using rules described by Zieky (1993, p. 342). On the SAT, category A items are preferred. The use of category B items is permitted. The use of C items is to be avoided. On the six operational forms administered in 1999, the African American/White DIF

analysis of 468 unique SAT-Verbal items resulted in 452 (96.6%) category A items, 15 (3.2%) category B items, and one category C item (0.2%). DIF screening on pretest items is very effective at screening out C items. On the one C item, African Americans performed better than matched Whites.

6. DIF/Difficulty Correlations

Most researchers would have concluded that DIF screening of the SAT had successfully screened items for DIF for African American test-takers on these two forms and stopped here. Santelices and Wilson (2010), however, looked to correlations to establish a case for bias that could not be made on the basis of DIF. Despite its emphasis on correlations, their study did not report all pertinent correlations. For example, the fact that the DIF variations were nearly perfectly correlated was not reported. Had these correlations been reported, it would have been clear that much of the analyses in their article would have been unnecessary. There is little reason to expect any meaningful differences in correlations between DIF and difficulty for different highly related standardization indices. This fact could be inferred simply from examination of the equations, as well as from the very high correlation among these DIF indices. Careful reading of Dorans (2004a) would have revealed that the criticisms were directed at overstated DIF effect sizes, an item-level phenomenon, and the R-score, not the correlation of DIF with difficulty.

In addition, Santelices and Wilson (2010) did not contain the high correlations between observed and estimated difficulty in the African American test-takers. The four measures of difficulty in Table 2 correlate .98 to 1.00, indicating that items are being similarly ordered by difficulty in the total, African American, and White test-takers. $P_{a\text{-}a}{}^{+}$ has its highest correlation (.99) with the Est $P_{a\text{-}a}{}^{+}$, produced by the standardization procedure. Est $P_{a\text{-}a}{}^{+}$ also has the closest mean to $P_{a\text{-}a}{}^{+}$, but the standard deviation of Est $P_{a\text{-}a}{}^{+}$ is closer to the standard deviations for the White and total test-takers.

Which of the correlations in Table 2 between difficulty and DIF is the most relevant? The correlation of -.18 between $P_{a\text{-}a}{}^{+}$ and DIF is the most pertinent correlation because it relates, across all items, DIF, the difference between $P_{a\text{-}a}{}^{+}$ and estimated $P_{a\text{-}a}{}^{+}$ in the African American test-takers, with $P_{a\text{-}a}{}^{+}$ in the same group of test-takers. About 3% of variation in the $P_{a\text{-}a}{}^{+}$ is shared with variation in DIF. In contrast, 98% of variation in $P_{a\text{-}a}{}^{+}$ is shared with variation in Est $P_{a\text{-}a}{}^{+}$

(which is based on the performance of White test-takers). These results are consistent with very little DIF.

Correlations between DIF and difficulty, the core of the Freedle (2003) and Santelices and Wilson (2010) bias claims, however, should be viewed with caution. These correlations vary with samples of people, and more importantly, items. DIF is an item-level procedure. Correlations across items pertain to that set of items, not any one item. A very easy and a very hard item that happen to exhibit nonzero DIF in opposite directions due to sampling variability can induce a nonzero correlation. In addition, item difficulty for the African American test-takers, $P_{a\text{-}a}{}^{+}$, and the difference between $P_{a\text{-}a}{}^{+}$ and Est $P_{a\text{-}a}{}^{+}$, which is STND DIF, both contain a common term, $P_{a\text{-}a}{}^{+}$. As explained in the appendix, it is reasonable to expect nonzero correlations between DIF and difficulty because they have $P_{a\text{-}a}{}^{+}$ in common.

In which direction should this nonzero correlation be? Hard items are less reliable for the lower scoring group. Consequently, the relationship between item performance and total score will be weaker for items on the hard half of the test to varying extents for both groups (see Table 4 of Dorans & Zeller, 2004b). It may be reasonable to expect overprediction for the African American test-takers on these items. Since the average DIF value is zero, any negative DIF on the hard items must be compensated for by positive DIF elsewhere. This suggests an expected negative correlation between item difficulty and DIF on the basis of item reliability considerations.

A more critical question is: How pertinent are these fluctuating correlations to fairness? Santelices and Wilson (2010) claimed that a negative correlation between DIF and difficulty (African American test-takers doing slightly better on hard items than White test-takers with scores comparable to their scores, and slightly less well on easy items) indicates test bias against African American test-takers. Would the test be biased in favor of African American test-takers if the correlation were positive? What does the direction of this correlation have to do with bias? It is the magnitude of DIF on an item that matters, not the correlation between DIF and difficulty.

In sum, Santelices and Wilson's (2010) analyses failed to find more than the usual amount of small DIF for African Americans on items from the two SAT editions they examined, which was expected, given extensive DIF screening prior to their use as scored items. Absent any sizeable DIF, the study leaned on correlations as evidence of bias. The correlations studied were

unstable and open to ambiguous interpretation. On one edition, the same set of items produced the correlations of -.41 and -.14. Santelices and Wilson focused on the correlation of -.41. Analysis on the complete data set for these items, however, resulted in a correlation of -.18 in the African American test-takers. (For completeness, on the other edition studied by Santelices and Wilson, the one administered in November 1994 to a full nationwide sample containing 222,098 White and 29,648 African American test-takers, the correlation between DIF and $P_{a\text{-}a}{}^{+}$ was -.02, close to zero.)

7. Unfair Treatment vs. Confirmation Bias

I have demonstrated that Santelices and Wilson (2010) misrepresented the number of SAT forms used in their study, misrepresented or misunderstood arguments made in critiques of Freedle's (2003) R-score, and changed "more unusual" into "considered serious," which created the false impression that the word *serious* was used by Dorans and Holland (1993). In addition, the SAT forms examined in their study were essentially DIF-free. Finally, I questioned the connection between DIF/difficulty correlations and fairness.

Santelices and Wilson (2010) contend that my criticism of Freedle (2003) was tantamount to arguing that the correlation between DIF and difficulty was an artifact due to choice of standardization index. Consequently, a study was designed and executed that demonstrated the obvious—highly correlated indices of DIF correlate with item difficulty to the same degree. I don't believe their study objectively addressed my concerns. Therefore, the conclusion by Santelices and Wilson (p. 127), "As independent researchers, we have objectively addressed the criticisms of Freedle and found that his findings still hold," might be viewed as another misrepresentation. (By independent, I infer unaffiliated directly with a testing company.)

In the conclusion section, Santelices and Wilson (2010, p. 127) state, "Tragically, the dismissal of his work has stopped involved and concerned parties from asking and discussing substantive, challenging questions about fairness in access to higher education." The article suggests that I treated Freedle (2003) unfairly. Unsubstantiated accusations of unfair treatment can have a tragic impact on the careers and lives of individuals.

In contrast to unfair treatment, which is an action committed by an individual that might only be excused under mitigating circumstances, confirmation bias is a part of the human condition. Our perceptions are affected by our expectations and opinions. They can be influenced

by emotional appeals that resonate within us. It is a challenge to keep confirmation biases, which provide solace in a complex, often confusing world, from being a barrier to understanding.

Near the end of a chapter titled "Illusions of Patterns and Patterns of Illusions" in *The Drunkard's Walk*, Mlodinow (2008, p. 189) discussed the pervasive problem of confirmation bias and quoted Francis Bacon's *Novum Organum*, which was published in 1620:

> The human understanding, once it has adopted an opinion, collects any instances that confirm it, and though the contrary instances may be more numerous and more weighty, it either does not notice them or else rejects them, in order that this opinion will remain unshaken.

Bacon remains relevant. Freedle (2003) claimed the SAT was culturally and statistically biased. Dorans (2004a) demonstrated that his claim of statistical bias was based on flawed calculations. Santelices and Wilson (2010) claimed to have addressed my objections but failed to do so. Instead an easy to refute hypothesis was attributed to me and easily refuted.

Mlodinow (2008, p. 190) added,

> To make matters worse, not only do we preferentially seek evidence to confirm our preconceived notions, but we also interpret ambiguous evidence in favor of our ideas. This can be a big problem because data are often ambiguous, so by ignoring some patterns and emphasizing others, our clever brains can reinforce their beliefs even in the absence of convincing data.

There was no differential item functioning on the forms studied by Santelices and Wilson (2010), as expected, given the extensive DIF screening. Instead, Santelices and Wilson resorted to ambiguous unstable correlations between DIF and difficulty as a measure of fairness, and justified the claim "…that SAT items do function differently for the African American and White subgroups in the verbal test…" (p. 106), when in fact they did not.

The evidence of confirmation bias in Santelices and Wilson (2010) is pervasive. The misattribution of a hypothesis about the effect of the DIF index employed on the correlation of DIF and difficulty to Dorans (2004a) is at the core of the evidence. Dorans made only a single passing reference to this correlation. The misattribution of *serious* to Dorans and Holland (1993) falsely inflated the severity of the DIF. Publication of misleading information about the number

of test editions studied inflated the very limited generalizability of the results. Omission of relevant correlations among DIF indices that may have led reviewers to question the merit of the study was also consistent with the hypothesis of confirmation bias.

Finally, the apparent belief that "independent" research is synonymous with objective research is also evidence of confirmation bias. It is consistent with the expectation that a researcher affiliated with a testing company cannot be objective. For 30 years, I have developed, researched, and used methods for assessing fairness in items, tests, and scores because fairness should be important to any professional. Proper understanding of the tools used to assess the fairness of assessment instruments and their uses is also important.

DIF, as noted earlier, is the wrong tool for assessing the fairness of test score use. Differential prediction is a more appropriate tool. But it has its limitations. The emphasis on DIF/difficulty correlations in the Freedle (2003) and Santelices and Wilson (2010) *Harvard Educational Review* articles brings to mind the classic quote from Kaplan (1964, p. 11) in his *Conduct of Inquiry*:

> There is a story of a drunkard searching under a lamp for his house key, which he dropped some distance away. Asked why he didn't look where he dropped it, he replied, "It's lighter here!" Much effort, not only in the logic of behavioral science, but in behavioral science itself, is vitiated, in my opinion, by the principle of the drunkard's search.

Looking at DIF/difficulty correlations is a drunkard's search when it comes to better understanding how test scores affect equity in higher education.

References

Bridgeman, B., & Burton, N. (2005, April). *Does scoring only the hard questions on the SAT make it fairer?* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Dorans, N. J. (2004a). Freedle's table 2: Fact or fiction. *Harvard Educational Review, 74*(1), 62–72.

Dorans, N. J. (2004b). Using population invariance to assess test score equity. *Journal of Educational Measurement, 41*(1), 43–68.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Liu, J. (2009). *Score equity assessment: Development of a prototype analysis using SAT Mathematics test data across several administrations* (ETS Research Rep. No. RR-09-08). Princeton, NJ: ETS.

Dorans, N. J., & Zeller, K. (2004a). *Examining Freedle's claims in his Harvard Educational Review article about bias and his proposed solution: Dated data, inappropriate measurement and incorrect and unfair scoring* (ETS Research Rep. No. RR-04-26). Princeton, NJ: ETS.

Dorans, N. J., & Zeller, K. (2004b). *Using score equity assessment to evaluate the equitability of a hard half test to a total test* (ETS Research Rep. No. RR-04-43). Princeton, NJ: ETS.

Freedle, R. O. (2003). Correcting the SAT's ethnic and social bias: A method for reestimating SAT scores. *Harvard Educational Review, 73,* 1–43.

Kaplan, A. (1964) *The conduct of inquiry: Methodology for behavioral science.* San Francisco: Chandler.

Kulick, E., & Hu, P. G. (1989). *Examining the relationship between differential item functioning and item difficulty* (College Board Rep. No. 89-5; ETS Research Rep. No. RR-89-18). New York, NY: College Entrance Examination Board.

Livingston, S. A., & Dorans, N. J. (2004). *Graphical item analysis* (ETS Research Rep. No. RR-04-10). Princeton, NJ: ETS.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential validity and prediction of the SAT* (College Board Research Rep. No. 2008-4). New York, NY: The College Board.

Mlodinow, L. (2008). *The drunkard's walk: How randomness rules our lives*. New York, NY: Pantheon Books.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models of culture-fair selection. *Journal of Educational Measurement, 13*, 3–29.

Sackett, P. R., Borneman, M., & Connelly, B. S. (2008). High stakes testing in education and employment: Evaluating common criticisms regarding validity and fairness. *American Psychologist, 63*, 215–227.

Santelices, M. V., & Wilson, M. (2010). Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review, 80*(1), 106–134.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zwick, R. (2002). *Fair game: The use of standardized admission tests in higher education.* New York, NY: RoutledgeFalmer.

Notes

[1] Freedle (2003) recommended using two scores for reporting the performance of African American test-takers on the SAT-Verbal: one score based on all the items and the other based on questions that composed the harder half of the SAT-Verbal exam (called the R-score).

[2] Some restrictive psychometric models predict a zero correlation between DIF and difficulty. These models make unrealistic assumptions about test-taker behavior. Test-takers who strive to maximize test performance do not conform to models that fail to account for how they behave when facing very hard questions.

[3] The sample sizes for the African American test-takers in Santelices and Wilson (2010) were not large enough to provide stable results for DIF analysis. A 78-item test has nearly 100 score points. With only 700 or 900 test-takers, there are many score points where the P^{+} is based on very small numbers. This leads to unstable estimates. In practice, smoothing is used to mitigate the effects of these small numbers (Livingston & Dorans, 2004).

Appendix

A proportion-correct P^+ can be expressed as a product of two vectors, a conditional proportions-correct vector, $\mathbf{p_g}$, which contains the proportions-correct at each score level for group g, and a weight vector, $\mathbf{w_g}$, which contains a weight for each score level, such that the sum of weights equals 1. For example, if there are 1,000 people in group g, and 2 at the highest possible score, then w for that highest score is 2/1,000 or .002. Let $\mathbf{w_g}$ and $\mathbf{p_g}$ be expressed as rows of numbers, and let $\mathbf{p'_g}$ be $\mathbf{p_g}$ expressed as a column of numbers.

Proportion-correct for group g is then $P_g^+ = \mathbf{w_g}\,\mathbf{p'_g}$. Here, there are several groups: the African American (a-a) test-takers, the White (w) test-takers, the Total (t) test-takers, and those test-takers neither African American nor White, the Other (o) test-takers. The following equations define the various P^+ referred to above: $P_{a\text{-}a}^+ = \mathbf{w_{a\text{-}a}}\,\mathbf{p'_{a\text{-}a}}$, $P_w^+ = \mathbf{w_w}\,\mathbf{p'_w}$, $P_o^+ = \mathbf{w_o}\,\mathbf{p'_o}$, and $P_t^+ = \mathbf{w_t}\,\mathbf{p'_t}$. Let $\mathbf{n'_g}$ contain the proportions of the Total group that come from the groups a-a, w, and o, and let $\mathbf{p^+_g}$ contain the various P^+. Then $P_t^+ = \mathbf{p^+_g}\,\mathbf{n'_g} = P_{a\text{-}a}^+ n_{a\text{-}a} + P_w^+ n_w + P_o^+ n_o$. Finally, STND P-DIF is the difference, $P_{a\text{-}a}^+ - \text{Est}(P_{a\text{-}a}^+) = \mathbf{w_{a\text{-}a}}\,\mathbf{p'_{a\text{-}a}} - \mathbf{w_{a\text{-}a}}\,\mathbf{p'_w}$. Note that both P_t^+ and STND P-DIF contain $\mathbf{w_{a\text{-}a}}$, $\mathbf{p'_{a\text{-}a}}$, and $\mathbf{p'_w}$. $P_{a\text{-}a}^+$ is the first part of STND P-DIF. Hence, STND P-DIF should be correlated with both P_t^+ and $P_{a\text{-}a}^+$ across items because they include common data, namely $P_{a\text{-}a}^+$. Likewise, P_t^+ is correlated with $P_{a\text{-}a}^+$ and P_w^+. There is little reason to expect a zero correlation between DIF and difficulty with real data, while DIF may be present in small but nonzero amounts.